

ARTIFICIAL INTELLIGENCE AND PEER REVIEW CAPACITY IN FORESTRY: A DECADE OF EVOLUTION AND AN INTRODUCTION TO PROPOSED MCFNS GUIDELINES FOR AI-ASSISTED REVIEW

AARON R. WEISKITTEL

Center for Research on Sustainable Forests, School of Forest Resources, University of Maine, Orono, ME 04469, USA

ABSTRACT. Ten years after an opinionated discussion of traditional peer-review processes and their alternatives was published in this journal, the landscape of scientific publishing in forestry has changed considerably. The fundamental supply-demand imbalance identified in 2015 has worsened: reviewer non-response rates already averaged 47.8% in 2015, and projection models now indicate a demand-to-supply ratio rising from roughly 1.0:1 in 2002 to 4.5:1 in 2024 and a projected 9.4:1 by 2032. Against this backdrop, artificial intelligence (AI) tools have moved from speculative alternatives to active components of many researchers' review workflows, with a 2025 survey reporting that 53% of researchers used AI assistance in peer review (Frontiers Media, 2025). Yet journal-level guidance has lagged adoption: most major forestry journal publishers (Elsevier, Springer Nature, Taylor & Francis) now bar reviewers from uploading manuscripts to AI tools, while Frontiers uses an AI assistant for research integrity assessments, leaving the active middle ground of AI-assisted review largely ungoverned. This paper revisits the four structural premises and four alternative review models discussed in 2015, assesses how each has evolved over the intervening decade, and introduces a new MCFNS section on AI applications in forestry. It also provides context for the accompanying AI-Assisted Review Guidelines, proposed for discussion and voluntary use, summarizes empirical findings on AI versus human review performance (including recent biotechnology-sector evaluation of GPT-5, Qwen-Plus, and Gemini 2.5 Pro across 398 open peer-reviewed preprints), situates the MCFNS approach within the shadow AI versus governed AI framework recently formalized in the operations literature, and presents an anonymized scored example applying the structured framework to a recently declined MCFNS submission. As a demonstration of the framework's applicability, this manuscript was self-assessed using the same B1–B12 rubric prior to publication; the evaluation, including pre- and post-revision scores and a full issue disposition table, is reported in Appendix A. MCFNS, which pioneered open and hybrid peer-review in forestry, is a natural venue to lead a second wave of reform with a structured, evidence-based framework for responsible AI integration.

Keywords: peer review, artificial intelligence, large language models, scientific publishing, forestry journals, MCFNS

1 INTRODUCTION AND RETROSPECTIVE ON WEISKITTEL (2015)

Peer-reviewed publications remain the primary mechanism for communicating scientific advances in forestry and natural-resource sciences. In 2015, I published an opinionated discussion of traditional peer-review processes and their alternatives in this journal (Weiskittel, 2015), documenting a system under serious structural strain. The analysis drew on personal experi-

ence as an author, reviewer, and associate editor for four international forestry journals, and on a sample of 64 manuscripts to quantify non-response and non-availability rates across four journals.

That paper organized its critique around four explicit premises that it argued were no longer reliably true for the forestry literature: (i) that a sufficient supply of willing and qualified reviewers exists relative to manuscript volume; (ii) that reviewers are adequately motivated by professional norms and reputational incentives alone;

(iii) that the process is substantively unbiased in its selection and evaluation of reviewers; and (iv) that double-blind review conducted by multiple independent reviewers provides adequate quality assurance. It then evaluated four structural alternative models: pre-publication review, post-publication review, high-volume journals accepting on scientific soundness rather than novelty, and independent professional review organizations.

Ten years on, each premise and each alternative model can be re-examined against a fuller record. This paper does so systematically in Section 2, then turns to AI as the new variable that was absent from the 2015 analysis but now reshapes every one of those four premises. The structure broadly mirrors the 2015 paper: background on the structural problem (Section 2), assessment of AI as both a cause of additional strain and a partial remedy (Sections 3 and 4), specific recommendations for stakeholders (Section 5), a worked example applying the proposed MCFNS framework to a recently declined submission (Section 6), and a self-assessment of this manuscript using the same B1–B12 scoring framework (Appendix A). Figures 1 and 2 illustrate the supply-demand trajectory and the AI versus human reviewer emphasis asymmetry that motivated the guidelines. The intent is the same as in 2015: to raise the importance of this issue and provide actionable suggestions for change.

2 A DECADE OF CHANGE: REASSESSING THE 2015 FRAMEWORK

2.1 Reviewer supply and demand: premise (i) has worsened

The supply-demand trajectory documented in 2015 has continued in the expected direction (Fig. 1). Russell and Weiskittel (2026) compiled SCImago-indexed publication counts for 156 forestry journals and detailed editorial metrics for 10 selected titles, and reported a more than fourfold increase in annual forestry articles from 3,519 in 2000 to 14,465 in 2024, with a 72% increase occurring in the nine years between 2015 and 2024. Reviewer non-response and non-availability rates, reported at $47.8 \pm 25.6\%$ in 2015 based on 64 manuscripts across four journals, are consistent with more recent anecdotal reports, though systematic data on reviewer response rates remain scarce. Russell and Weiskittel (2026) indexed forestry article output and the US forestry researcher workforce to a 2002 baseline and, after accounting for projected workforce reductions following 2025 US federal research cuts, estimated that the demand-to-supply ratio rises from 1.0:1 in 2002 to 4.5:1 in 2024 and a projected 9.4:1 by 2032. The projection carries uncertainty, but the direction is consistent with the experience

of associate editors across multiple forestry journals over the past decade.

Contributing factors include continued graduate enrollment declines in traditional forestry programs in North America and Western Europe, ongoing contraction of USFS intramural research capacity, and a growing share of submissions from the Global South where the qualified reviewer pool does not yet proportionally match submission volume. Geographic representation gaps between submission pools and editorial boards are well documented across the life sciences (Tennant et al., 2016), and the pattern applies to forestry journals: submission volumes increasingly originate from Asia and the Global South while editorial board composition has not diversified at the same rate, a disparity that affects both review availability and the cultural calibration of editorial standards.

2.2 Reviewer incentives: premise (ii) shows marginal gains

The incentive structure for reviewers identified by Weiskittel (2015) has improved in form but not in outcomes. Publons, which was cited in 2015 as a promising emerging platform for credentialing review contributions, was acquired by Clarivate in 2017 and integrated into Web of Science profiles in 2019. Reviewer recognition is now more visible, and some publishers issue formal certificates and award fee waivers. The reviewer exchange model, in which the act of publication requires contributing a review, has been piloted by several society journals. In principle this directly aligns the incentive to submit with an obligation to review; in practice, the approach is undermined by the near-limitless number of publication outlets now available, which allows authors to route submissions to journals that impose no such requirement. Despite these developments, institutional evaluation committees at universities and research agencies still measure productivity primarily through publications and external funding. Completed peer reviews remain functionally invisible in formal promotion and tenure assessments regardless of their quality or quantity. This is the principal-agent problem (García et al., 2015) that no credentialing platform has yet resolved. The incentive to review has not improved proportionally to the incentive to submit.

2.3 Alternative review models: a ten-year accounting

The 2015 paper evaluated four structural alternatives to traditional peer review. Their trajectories over the past decade are instructive. Pre-publication review through preprint servers has grown considerably: ESSOAr (earth and space sciences) launched in 2018 and EarthArXiv in

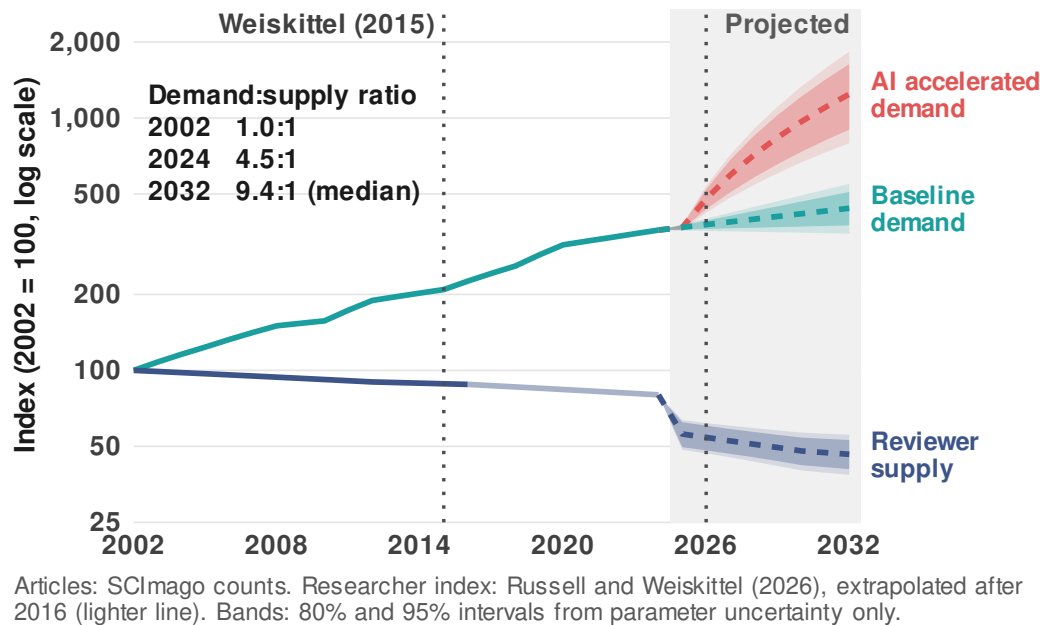


Figure 1: Reviewer supply and demand trajectory in forestry journals. Index (2002 = 100) of forestry article output, observed to 2024, and the US forestry researcher workforce, observed to 2016 and extrapolated to 2024 by Russell and Weiskittel (2026), with projections to 2032 that follow their assumptions. Solid lines show 2002 to 2024 values; dashed lines show the median of a Monte Carlo projection ($n = 5,000$ iterations), with shaded bands giving the 80% (darker) and 95% (lighter) intervals from parameter uncertainty only. Parameter distributions: annual growth rate for baseline demand $\sim \text{Normal}(0.025, 0.015)$, centered on their 2.5% rate; terminal 2040 multiplier for the AI accelerated scenario $\sim \text{Uniform}(3, 7)$, ramped linearly from 2025 and consistent with the range of marginal cost effects discussed by Cunningham (2026); reviewer pool cut in 2025 $\sim \text{Uniform}(0.20, 0.40)$ and annual decline $\sim \text{Uniform}(0.015, 0.045)$ to 2030, halved thereafter, centered on the 30% cut and 3% decline in their projection. Dotted vertical lines mark the 2015 framework paper and the present year.

2017, while biorxiv (launched 2013) demonstrated a viable model for biological sciences. Adoption in forestry and natural-resource sciences remains notably lower than in ecology or atmospheric sciences, and preprints have not displaced formal peer review as the primary quality-assurance mechanism in forestry publication.

Post-publication review through platforms such as PubPeer and F1000Research has also expanded, but has functioned primarily as an error-correction mechanism rather than a gatekeeping alternative. The high-volume soundness-only model, represented most prominently by PLOS ONE, has influenced journal editorial policy across several publishers; a number of journals now distinguish explicitly between novelty and soundness in their editorial criteria, which is a substantive shift from 2015 practice. Independent professional review organizations have fared less well: Rubriq, which the 2015 paper noted as an emerging option, ceased operations in 2019; Axios Review closed in 2017; Peerage of Science persists but remains a niche platform. The overall pattern is that structural alternatives to traditional

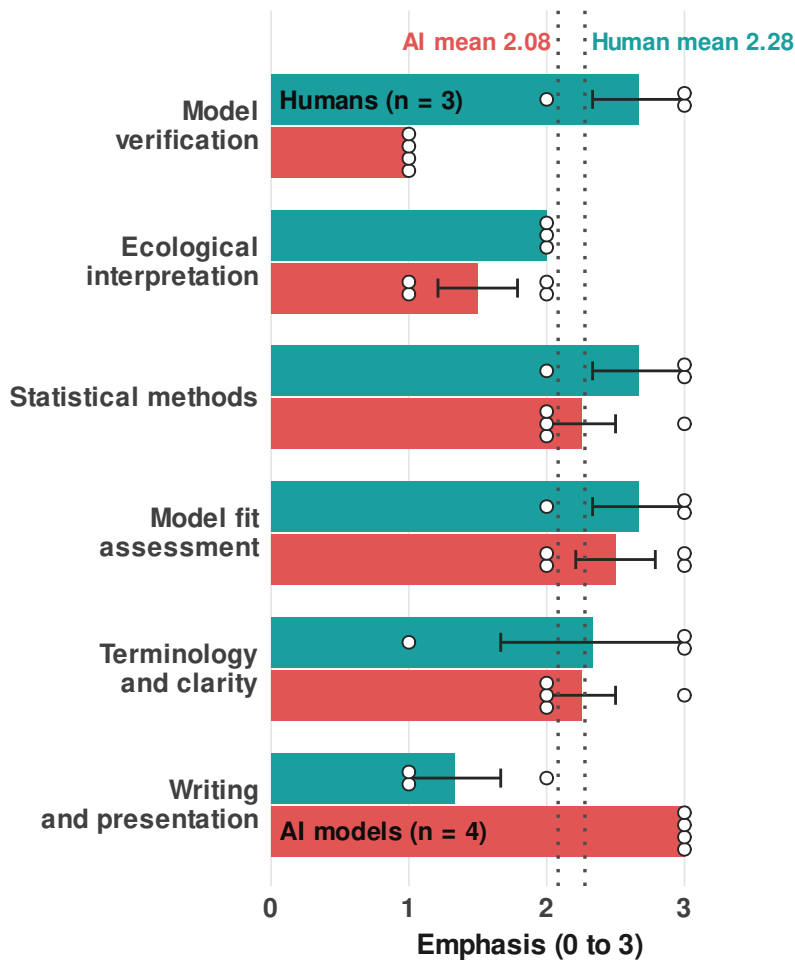
review have expanded the ecosystem without replacing the core model.

2.4 Bias and double-blind review: premises (iii) and (iv) revisited

The 2015 paper noted that double-blind review has limited practical effectiveness in small, specialized fields such as forestry because author identity is often inferrable from the citation network, study area, and methodological approach alone. This limitation has likely deepened. AI-powered stylometric tools have been shown to identify probable authorship from prose patterns with considerably greater reliability than manual inference when trained on an author's prior publication record (Koppel et al., 2009), raising the possibility that the anonymization double-blind review depends on is more permeable than it was a decade ago. The field has not yet formally assessed this risk, and it warrants attention before AI stylometric tools become routine in editorial workflows.

On the question of process bias, the 2015 paper raised concerns about systemic disparities in review assignment

A. Review emphasis



Mean ± 1 SE (sample SD); no bar where all raters agreed.

B. Worked example scores

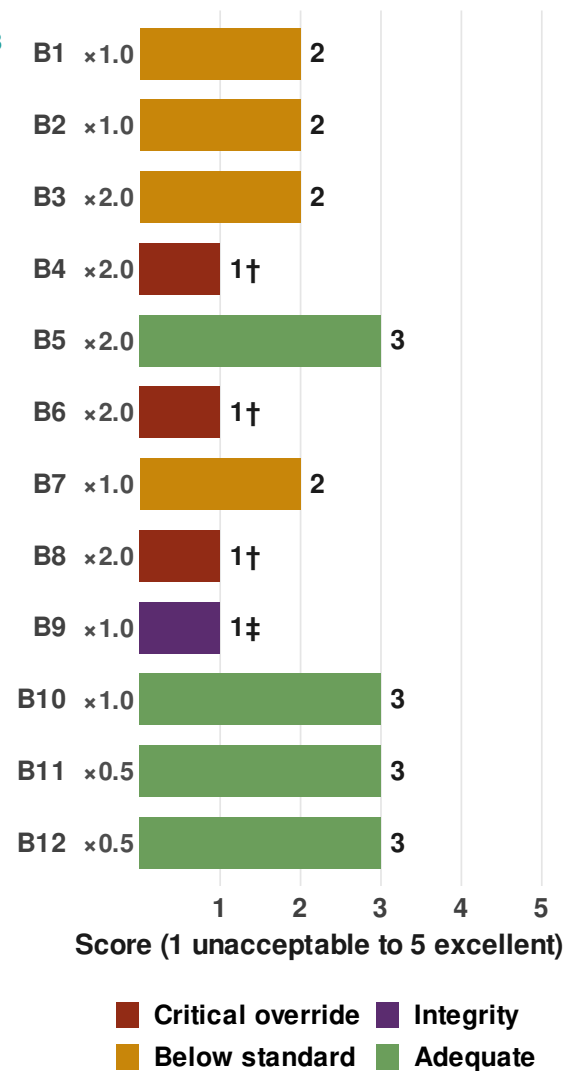


Figure 2: (A) Emphasis on six review categories by three human reviewers (teal) and four large language models (red) reviewing the same manuscript, scored from 0 (not mentioned) to 3 (major focus) by Russell and Weiskittel (2026). Bars show means with ± 1 SE, points show individual reviews, and dotted vertical lines mark the AI mean (2.08) and human mean (2.28). The two largest divergences are Model verification (humans exceed AI by 1.67 points) and Writing and presentation (AI exceeds humans by 1.67 points). (B) Anonymized scoring matrix from the worked example declined submission, showing all twelve MCFNS evaluation dimensions (B1 through B12) with dimension weight, score (1–5), and critical-override flags (†). Weighted total: 29/80 (36%). A score of 1 on any $\times 2$ dimension triggers automatic Reject/Major Revision regardless of aggregate, and † marks the Level 3 citation integrity finding on B9, which requires rejection under the escalation pathway independently of the score.

and outcomes related to author geography and institutional affiliation. The pattern noted in Section 2.1, in which editorial board composition has not diversified proportionally to submission pools, suggests those concerns remain relevant. AI-assisted manuscript screening may either compound this bias if models are trained on historically skewed acceptance patterns, or reduce it if

structured checklists derived from scientific criteria alone are applied consistently across submissions regardless of author affiliation. Which effect dominates depends on how the tools are deployed and whether their application is monitored against demographic outcomes. The COPE guidelines (COPE, 2019) provide a framework for

editors considering bias mitigation in review assignment, independent of AI adoption.

2.5 AI as a response to structural strain and a new source of it

The adoption of AI tools in research workflows emerged partly as a rational response to mounting review burden, as Kustov (2026) documents in a widely cited analysis of AI’s disruption of academic practices. Researchers facing non-response from invited reviewers and editors managing large queues with inadequate associate editor capacity turned to available tools to improve efficiency. The same dynamic, however, creates a new source of strain. If AI considerably reduces the marginal cost of manuscript preparation (Cunningham, 2026), submission volumes could increase fivefold even while journal publication slots remain fixed. Desk rejection rates would rise from roughly 50% to 90%. Peer review, already operating beyond sustainable capacity, becomes impossible at scale under those conditions without structural adaptation.

Russell and Weiskittel (2026) documented how publisher-level positions on AI in forestry peer review diverge: Elsevier, Springer Nature, and Taylor & Francis, which publish many of the forestry titles indexed by SCImago, now tell reviewers not to upload manuscripts to generative AI tools or to use AI to draft review text, while Frontiers journals use an AI assistant for research integrity checks. Journal-level guidance between those positions remains sparse, and the operative regime for most forestry reviewers is therefore unsupervised individual use of tools their publisher formally forbids. The gap between evolving practice and available guidance is the operational problem that the MCFNS guidelines address. Russell and Weiskittel (2026) explicitly point to the MCFNS approach described here as a candidate example of structured, journal-level reviewer guidance that could narrow that gap.

A complementary pressure operates at the system level. Cieszewski (2026), in this issue, argues that the sheer volume of publishing, independent of any particular AI tool, functions as a more corrosive threat than AI capability itself. Large commercial publishers, supported by infrastructure for rapid processing, attract a rising share of submissions and a self-reinforcing increase in volume. The downstream consequence is the proliferation of papers whose own authors do not cite them in subsequent work, which erodes both the citation ecosystem and the signal value of publication. AI use in writing and review is usefully analogous to earlier methodological shifts such as stepwise regression and mixed-effects modeling: each enabled the effortless production of complex results that may appear correct on inspection, but

where the earlier tools were specific and conspicuous, AI is flexible enough that its use is easily concealed. The operational risk confronting forestry journals is not AI capability in isolation but the combination of rising submission volume with opaque AI assistance on both the authoring and reviewing sides of the process.

3 WHAT AI DOES AND DOES NOT DO WELL

3.1 The human and AI review asymmetry

In a structured case study, a single forestry manuscript (Chen et al., 2023) was independently reviewed by three human reviewers (the handling associate editor and two referees) and four large language models (ChatGPT GPT-5.2, Claude Opus 4.5, Gemini Pro v3, and Perplexity), and each review was scored for emphasis on six categories from 0 (not mentioned) to 3 (major focus) (Russell and Weiskittel, 2026). Human reviewers and AI placed similar emphasis on model fit assessment (2.67 versus 2.50) and on terminology and clarity (2.33 versus 2.25), with overall mean scores of 2.28 (humans) and 2.08 (AI). The categories with the largest divergence showed AI placing less emphasis on model verification (human mean exceeded AI by 1.67 points on the 0–3 scale) and more on writing and presentation (AI mean exceeded human by 1.67). Human reviewers also placed more emphasis on statistical methods (2.67 versus 2.25) and ecological interpretation (2.00 versus 1.50), and all three asked for tests of model performance by species, region, or diameter class, which the AI models mentioned only briefly (Fig. 2A).

This asymmetry is empirically documented, not a philosophical claim about machine cognition. Kustov (2026) frames the same point usefully: the productive concerns about AI tools are about verification and security, not whether large language models achieve some higher-order form of understanding. For the purposes of peer review, the relevant question is whether AI can reliably execute specific, well-defined evaluation tasks. The evidence indicates yes for a defined subset, and no for another. The MCFNS guidelines are designed to exploit this asymmetry by directing AI to the tasks where its performance is reliable and preserving human judgment for those where it is not.

Independent evidence from an adjacent field is consistent with the forestry-specific findings. Li et al. (2026) evaluated GPT-5, Qwen-Plus, and Gemini 2.5 Pro on 763 biotechnology preprints (398 with open peer reviews) and 12 NSF grant proposals. LLM-generated reviews produced substantive and well-structured comments with a strong emphasis on experimental design and statistical analysis, but were systematically more

lenient than human reviewers. On grant proposals, the three models clustered at 3.2–3.8 on a 1–4 scale while the human mean was 2.5; AI reviewers were 34% less likely than humans to comment on positioning and also asked for more citations. AI detectors failed to reliably identify AI-generated text in review comments, and simple rewording bypassed them. The authors conclude that LLMs can serve as useful ad hoc reviewers for screening and high-volume submissions but not as substitutes for human judgment, which is structurally the same recommendation as the one in Section 3.1 for forestry.

3.2 Citation integrity: a newly documented failure pattern

The supplementary worked example of the companion study (Russell and Weiskittel, 2026) characterized a citation integrity failure pattern not previously documented in the forestry peer review literature. Pattern A misattribution, a non-existent reference detectable by DOI resolution failure, was already known. Pattern B misattribution is newer and more difficult to detect: the model generates a citation to a real, existing paper (correct authors, journal, and DOI) whose actual content bears no substantive relationship to the claim for which it is cited. Detection requires reading the cited paper; DOI resolution cannot identify the error.

Reported citation hallucination rates vary widely across AI model vintages. Walters and Wilder (2023) found that 55% of GPT-3.5 bibliographic citations and 18% of GPT-4 bibliographic citations were fabricated in a controlled evaluation on scholarly topics, with additional substantive errors in citations that did resolve, and Russell and Weiskittel (2026) cite the same study as evidence that fabricated citations remain a documented harm of AI misuse in research. Kustov (2026) makes the uncomfortable observation that applying intense scrutiny to AI citation errors while tolerating a system in which informal estimates suggest that only about 4% of published papers are of good quality (Duijm, 2024) applies a double standard that does not withstand examination. The productive response is structured verification protocols, not categorical rejection of AI assistance. This perspective is operationalized in Module 4 of the accompanying prompt library.

3.3 Vocabulary and structural signals of AI-generated text

AI-generated academic text carries identifiable vocabulary and structural signatures. High-frequency lexical signals include terms such as: *delve*, *multifaceted*, *nuanced*, *pivotal*, *paramount*, *tapestry*, *seamless*, *harness*, *elucidate*, *underscores*, *testament*, *illuminate*. Structural signals include uniform paragraph length across

sections, conclusions that closely paraphrase the introduction, verbatim phrase repetition across non-adjacent sections, and systematic capitalization of common disciplinary terms as though they are proper nouns. Module 3 of the prompt library provides a standardized approach for screening these signals.

Two caveats are essential. First, smooth and polished prose is not itself an AI signal; many accomplished human writers produce fluent text, and penalizing clarity would be counterproductive. Second, non-native English writing patterns are explicitly not AI signals and must not be treated as such. The equity rule built into the Module 3 prompts prevents penalizing authors for language characteristics that have no bearing on scientific merit. Crucially, all Module 3 outputs are probabilistic screens that require human judgment; the module produces a risk level, not a finding.

3.4 Shadow AI and the case for a governed alternative

Bhargava et al. (2026) formalize a distinction that clarifies the operational stakes of the current moment. Under *shadow AI*, individual reviewers use general-purpose tools in an opaque and uncoordinated manner: the journal cannot observe which tools were used, cannot audit outputs, and crucially cannot measure the correlation induced across reviewer reports when multiple reviewers privately consult similar models. Under *governed AI*, the journal provides a scoped tool whose outputs are structured, logged, and auditable, and the correlation between the AI signal and human signals is an observable process parameter.

Their analytical model incorporates three forces simultaneously: AI tools increasing author productivity (submission rate λ); AI review quality (parameter θ); and the anchoring correlation ρ_{HA} between the AI report and human reviewer signals. Under correct generalized least squares aggregation, adding an AI signal cannot worsen posterior precision, and strictly improves it when ρ_{HA} stays below a redundancy threshold $\rho_{\text{red}} = \sigma_H/(m\sigma_A)$. The actionable implication is that the risk is not AI use in principle but the combination of private AI use across multiple reviewers, which induces hidden correlation among reviewer signals, together with the absence of a journal-provided baseline that would allow that correlation to be measured. Bhargava et al. also catalogue five operational concerns that any structured AI integration must address: reviewer complacency, reviewer over-correction, author gaming of the AI rubric, limited improvement value, and prompt injection. Each is consistent with the risks the MCFNS prompt library is designed to mitigate through auditable output formats and disclosure.

4 OVERVIEW OF THE PROPOSED MCFNS GUIDELINES

The accompanying Proposed MCFNS Guidelines for AI-Assisted Peer Review (2026 Edition, v1.0), available with this paper as a supplementary file, are a proposal for discussion and voluntary use, not journal policy. They are organized around four operational components, each directly motivated by an empirical finding or an editorial failure mode documented above.

The first component is a conceptual framework suggesting appropriate and inappropriate uses. AI assistance is welcome within a structured, human-in-the-loop framework. Manuscript content should not be uploaded to AI tools that store or transmit it externally without confirmed zero-data-retention agreements. Disclosure of AI use in a review is encouraged but not required.

The second component is a two-document review architecture that separates the internal EIC Report from the Author Decision Letter. These are not the same document in two different tones; they are structurally different instruments with different content rules and different audiences. Integrity concerns, AI generation hypotheses, COPE escalation language, and numerical scoring totals belong in the EIC Report and should not appear in the Author Decision Letter. Conflating the two documents exposes the journal to ethical and legal risk and violates COPE norms.

The third component is a six-module LLM prompt library (Modules 0–6b) ensuring that reviewers using different AI platforms produce structured outputs in a consistent, auditable format. Module 0 establishes behavioral constraints that prevent citation fabrication. Modules 1–5 cover preliminary screening, statistical rigor audit, AI content detection, citation integrity audit, and scoring scaffold, respectively. Modules 6a–6b produce EIC Report and Author Decision Letter drafts, with Module 6b containing built-in content exclusion rules and a pre-transmission check for excluded terminology.

The fourth component is a set of domain-specific method checklists for five forestry method families: growth and yield modeling; FIA and forest inventory; LiDAR and remote sensing; machine learning applications; and coupling coordination degree and multi-criteria index studies. These checklists address the most common manuscript-specific technical errors that generic statistical audits do not capture.

Positioned against the Bhargava et al. (2026) framework, the MCFNS guidelines occupy an intermediate posture between shadow AI and fully governed AI. MCFNS does not currently provide a dedicated journal-tuned model; that would be the fully governed case and would require infrastructure a small open-access journal is not positioned to operate. What the guidelines

do provide is a structured interface, consisting of the six-module prompt library with auditable outputs, that reviewers using any major AI platform can apply consistently. The operational content of Bhargava et al.’s recommendation to “scope AI to tasks where criteria can be specified and outputs can be validated” and to “prevent scope creep into value judgments” is directly reflected in the module structure: Modules 1–4 are scoped to auditable diagnostics, while Module 5 (scoring scaffold) and Module 6 (report generation) produce inputs to editorial judgment rather than substitutes for it. Whether MCFNS should pilot a journal-tuned AI reviewer in a future guidelines version remains an open question.

One check merits explicit attention beyond the general module structure: Module 2 check S2b flags manuscripts reporting Pearson correlation or ANOVA across spatially autocorrelated grid cells without autocorrelation correction. When Global Moran’s $I > 0.3$ and standard parametric tests are used on grid-cell data, the effective degrees of freedom may be 10–50× smaller than the nominal cell count, rendering tests anti-conservative. Required corrections include Dutilleul’s (1993) modified t -test (`R: SpatialPack::modified.ttest()`; Osorio and Vallejos 2024) or spatial lag and error regression (`spdep`, `spatialreg`). This check was developed from the worked example in Section 6 and represents a class of error that is common in grid-scale ecological and forestry manuscripts but is rarely caught by standard statistical checklists.

5 RECOMMENDATIONS

5.1 For authors

Authors are encouraged to disclose AI assistance, including which tools were used, for which tasks, and how outputs were verified before use. If 53% of researchers report using AI in review (Frontiers Media, 2025), the same proportion almost certainly uses it in writing, and the lexical and structural patterns are detectable by the tools described in Section 3.3. Authors in low-resource settings who rely on AI for language assistance are not the target of this disclosure recommendation; the concern is with undisclosed AI contributions to the scientific content and reference list. Stigmatizing AI assistance broadly would disproportionately harm researchers for whom it provides the greatest productivity benefit, which is inconsistent with the field’s stated commitment to international participation in science.

Treat reference lists with particular caution regardless of how they were generated. Cite papers you have read. Pattern B misattribution, real papers whose content does not match the cited claim, is one of the most reproducible failure modes of current AI models and is

the class of error that a reviewer reading the cited work will catch with high probability.

5.2 For reviewers

Use the structured prompt library rather than posing open-ended queries to an AI tool. Unstructured queries produce outputs that cannot be compared across reviewers, are biased toward superficial manuscript features, and cannot be audited. Structured module outputs, with mandatory location citations and confidence flags, can be compared, criticized, and revised. That auditability is what makes AI-assisted review compatible with scientific norms rather than in tension with them.

Apply independent expert judgment to all ecological validity, scientific significance, and model validation assessments. These dimensions fall outside the reliable performance range of current AI models, as suggested by the human and AI review asymmetry evidence reviewed in Section 3.1. The asymmetry principle is not a mechanism for reducing reviewer effort; it concentrates reviewer effort on the dimensions where domain expertise adds the most value and where AI assistance is least reliable.

Consider disclosing AI use in review reports in which it was employed. The proposed MCFNS guidelines encourage such disclosure, which is consistent with COPE expectations for transparent review processes.

5.3 For associate editors and the editor-in-chief

Maintain the two-document architecture without exception. Integrity concerns, escalation language, and scoring totals remain in the EIC Report. The Author Decision Letter communicates scientific findings and actionable revision requirements, not editorial suspicion.

Apply the three-level citation integrity escalation pathway when reviewer reports document citation concerns. Level 1 is a suspicion flag that requires requesting disclosure in the revision cover letter. Level 2 requires independent editorial verification of additional references; if two or more further misattributions are confirmed, escalation to Level 3 is required. Level 3 requires COPE referral and rejection on integrity grounds, with the author letter communicating the decision without integrity framing.

Keep the proposed guidelines open for discussion and update them as AI tools evolve. These guidelines are offered as v1.0, a proposal for discussion; they will require revision as the field accumulates experience with structured AI-assisted review. The versioning convention (v1.x) allows updates to be tracked and cited precisely.

5.4 For MCFNS as a journal: operational posture

Consistent with editorial discussion during the drafting of these guidelines, the suggestions in this paper are offered as hypothetical possibilities rather than firm policy prescriptions. The time and effort required for rigorous peer review has already become a scarce resource, and any proposal that adds reviewer workload without removing an equivalent load elsewhere is unlikely to survive contact with operational reality. MCFNS should therefore treat the prompt library as a voluntary resource offered to reviewers who wish to use AI assistance, rather than a required step in the review workflow. A modest intermediate adoption path is to encourage consistent use of Module 0 behavioral constraints and Module 3 screening outputs across reviewers who are already using general-purpose AI tools privately, which converts informal shadow AI activity into observable, auditable inputs without imposing new mandatory steps on volunteers.

A separate question raised during those discussions concerns whether the same tooling used by reviewers could be offered to authors for self-checking prior to submission. This is a genuine conundrum: some authors will use such tools to strengthen weak manuscripts in ways that reduce reviewer burden, while others will use them adversarially to disguise manuscripts that ought not to be submitted in their current form. The current guidelines do not resolve this tension. A reasonable posture for MCFNS is to make the prompt library publicly available so that authors who wish to pre-screen their own work can do so, while recognizing that some fraction will use the same tools to optimize against the journal's rubric rather than to improve the underlying science. The structural countermeasure is not to restrict author access to screening prompts, which is practically unenforceable, but to ensure that Module 3 structural screening and Module 4 citation integrity audits are applied systematically at review time and to vary rubric weightings over time so that pre-optimization cannot converge on a static target.

6 WORKED EXAMPLE: ANONYMIZED SCORING OF A DECLINED SUBMISSION

6.1 Background

The following presents the scored evaluation of a manuscript submitted to MCFNS in early 2026. All identifying details have been anonymized. The manuscript presented a fine-grained grid-scale coupling coordination degree (CCD) analysis of forest ecosystem resilience and recreational efficiency, integrating intangible cultural heritage (ICH) kernel density estimation as

a novel spatial indicator. The study area was approximately 1,800 km²; methods included entropy weighting, CCD modeling, kernel density estimation (KDE), InVEST Habitat Quality, Global Moran's *I*/LISA, and a spatially explicit obstacle model. The submission falls within the domain of quantitative, methodology-driven spatial forestry that MCFNS regularly publishes. As a transparency measure, this manuscript was also evaluated using the same framework; the screening results, scoring matrix, and revision decisions are documented in Appendix A.

6.2 Scoring matrix

Table 1 presents the full scoring matrix across the twelve MCFNS evaluation dimensions. The scoring scale (1–5) and weighted total structure are defined in detail in Annex A of the accompanying guidelines document. Dimensions marked with a critical override flag received a score of 1 on a $\times 2$ dimension, triggering automatic Reject/Major Revision regardless of the aggregate score. The overall weighted total of 29/80 (36%) falls below the 45% threshold required for Major Revision consideration, and three critical overrides (B4, B6, and B8) were triggered independently before the aggregate was calculated. Citation quality (B9) also scored 1, yet B9 carries a $\times 1$ weight, so it does not trigger the override. Its Level 3 citation integrity finding requires rejection on its own under the escalation pathway (Section 6.3).

Table 1: Anonymized scoring matrix for declined submission. Scale: 5 = Excellent, 4 = Good, 3 = Adequate, 2 = Below standard, 1 = Unacceptable. Wt. = dimension weight; † = critical override triggered (score of 1 on $\times 2$ dimension); ‡ = Level 3 citation integrity finding. Weighted total: 29/80 (36%).

Dim.	Evaluation dimension	Wt.	Score
B1	Originality and novelty	$\times 1$	2
B2	Scientific significance	$\times 1$	2
B3	Methodological rigor	$\times 2$	2
B4	Statistical appropriateness	$\times 2$	1 [†]
B5	Mathematical/model rigor	$\times 2$	3
B6	Model validation	$\times 2$	1 [†]
B7	Data availability	$\times 1$	2
B8	Code/reproducibility	$\times 2$	1 [†]
B9	Literature/citation quality	$\times 1$	1 [‡]
B10	Interpretation	$\times 1$	3
B11	Writing clarity	$\times 0.5$	3
B12	Figures and tables	$\times 0.5$	3
Weighted total			29/80 (36%)

† Critical override: score of 1 on any $\times 2$ dimension triggers Reject/Major Revision regardless of aggregate. Threshold: $\geq 45\%$ for Major Revision consideration. ‡ Level 3 citation integrity finding ($>20\%$ confirmed Pattern B misattribution), which requires COPE referral and rejection on integrity grounds independently of the score.

6.3 Key findings from the structured review

Citation integrity (B9, Level 3 integrity finding). Manual content verification of 35 cited references identified systematic Pattern B misattribution in approximately 11 cases, a rate of 31%. In each instance, the cited paper resolved correctly via DOI but its actual content bore no substantive relationship to the claim for which it was cited; the misattribution is at the level of content, not metadata. The escalation pathway (Level 3: COPE referral) was applied given the rate exceeded the 20% reporting threshold established in the guidelines.

Statistical validity (B4, critical override). The manuscript reported Pearson correlation and one-way ANOVA computed across spatially autocorrelated grid cells, while separately reporting a Global Moran's *I* that confirmed severe spatial autocorrelation in the primary response variable. Standard test statistics are anti-conservative under these conditions; the appropriate corrections (Dutilleul, 1993) were not applied.

Reproducibility (B8, critical override). The KDE bandwidth for the intangible cultural heritage density indicator, the manuscript's most novel contribution, was not stated anywhere in the methods. InVEST Habitat Quality threat parameters were absent. No code was archived. The central spatial agglomeration finding depends on an undisclosed bandwidth parameter and cannot be independently reproduced.

Validation (B6, critical override). No independent validation was conducted. The InVEST Habitat Quality model was applied without parameter documentation, sensitivity analysis, or comparison to any benchmark. The equal-weight assumption in the CCD model controlled zone characterization results without any sensitivity analysis across alternative weight ratios.

6.4 Anonymized excerpt from the Author Decision Letter

The following excerpt illustrates the two-document architecture in practice. The EIC Report documented citation integrity at Level 3 and recommended COPE referral. The Author Decision Letter communicated only the scientific findings, with no integrity framing:

*"The reviewer identified a number of instances in which cited references do not appear to support the specific claims made for them. Six representative examples are provided in the table below. We require a fully verified corrected reference list with a point-by-point response addressing each listed discrepancy. [...] The manuscript reports a Pearson correlation of $r = -0.62$ and one-way ANOVA computed across grid cells, while also reporting a Global Moran's *I* of 0.652 for the primary response variable. Standard correlation and ANOVA test*

statistics are anti-conservative under conditions of spatial autocorrelation of this magnitude. [...] The bandwidth used to compute the [KDE-based] indicator was not stated in the Methods section, making independent reproduction of this central result impossible."

No integrity framing, AI generation language, COPE terminology, or numerical scoring appeared in the author letter. Those details were documented in the EIC Report and handled through the appropriate editorial channel.

7 CONCLUSIONS

Ten years after Weiskittel (2015) identified serious strain in the forestry peer-review system, it is possible to offer a structured retrospective against the framework that paper proposed. Of the four premises examined in 2015, not one has improved materially. The first, adequate reviewer supply, has deteriorated at the rate the 2015 analysis projected and at the rate that a simple comparison of article volume growth to editorial board growth would predict. The second, adequate reviewer motivation, has seen credentialing infrastructure expand without producing measurable improvement in review availability or completion rates. The third, process unbiasedness, remains compromised by the geographic representation gap between submission pools and editorial boards, and is now further complicated by AI-assisted authorship detection capabilities that may systematically disadvantage authors whose stylistic patterns differ from dominant training corpora. The fourth, double-blind review as a quality assurance mechanism, faces a new threat in the form of AI stylometric identification that scales the author-inference problem the 2015 paper noted as already limiting in small specialized fields.

The four alternative models evaluated in 2015 have had uneven trajectories. Pre-publication preprints and high-volume soundness-only review have both grown and influenced practice, though neither has displaced traditional review as the primary mechanism in forestry. Independent professional review organizations have largely not survived commercially. Post-publication review has found a role in error correction but not in gatekeeping. Structural alternatives to traditional peer review remain insufficient on their own; the problem requires operational improvements to the existing process alongside any structural reform.

AI represents the new variable that the 2015 analysis could not anticipate. It is simultaneously a source of additional submission pressure and a practical tool for improving review throughput and consistency. The appropriate response is neither uncritical adoption nor reflexive rejection. It is structured integration, offered

initially as a set of hypothetical procedures rather than mandatory policy: clear suggested uses, voluntary assistance and disclosure, standardized prompt modules that produce auditable outputs, and domain-specific checklists that go beyond generic statistical guidance. Converging evidence from the governed AI framework of Bhargava et al. (2026) and the empirical evaluation of Li et al. (2026) suggests that this posture, which is intermediate between shadow AI and a fully journal-tuned model, captures most of the measurable benefit while remaining feasible for a journal at the scale of MCFNS. Appendix A demonstrates this in practice: the B1–B12 framework was applied to this manuscript prior to publication, producing eight specific findings that were addressed in revision and raising the weighted score from 73% to 77.5%, within the Minor Revision band. A system that was already failing by its own metrics before AI entered the picture cannot afford to respond to a new and substantially disruptive variable with the same institutional inertia that characterized the previous decade.

MCFNS established hybrid peer-review in forestry in 2009 (Cieszewski and Strub, 2009). The accompanying guidelines continue that tradition of proactive reform. The specific invitation to other journals in the natural-resource sciences is fourfold: open a formal discussion of AI use in review before practice outpaces it further, not after; publish any resulting guidelines with a version number so that updates can be cited and compared across venues; apply domain-specific method checklists rather than generic statistical audits, because the most consequential errors in forestry manuscripts are domain errors; and contribute calibration experience back to the community, because the B1–B12 rubric will improve only as evidence from its application accumulates. The goal is not uniformity of policy but interoperability of process: a peer-review system for forestry that can identify its own failure modes and correct them across journal boundaries, rather than repeating the decade of institutional inertia that the 2015 paper documented and that this one confirms has continued.

Comments on the proposed guidelines, and on the questions raised in this paper, are invited. Readers, reviewers, authors, and editors can submit them to MCFNS, including as discussion papers. The guidelines will remain at the proposal stage during this period of consideration and discussion, and their next version will build on the feedback received.

ACKNOWLEDGMENTS

Thanks to Chris J. Cieszewski for his leadership of MCFNS and openness to proactive policy development, and to Mike Strub for comments on the first draft. Matt Russell contributed to the empirical analysis underly-

ing the companion Forest Ecosystems manuscript. The worked example in Section 6 was handled in the author's capacity as manuscript editor for MCFNS. Funding was provided by the University of Maine Center for Research on Sustainable Forests and the NSF EPSCoR Maine-FOREST E-RISE program (Award No. OIA-2416918). The author used Claude (Anthropic, claude-sonnet-4-6) for initial structural drafting and prose editing of this manuscript; all content was critically reviewed, substantially revised, and verified by the author prior to submission, in line with the disclosure encouraged in the accompanying guidelines.

REFERENCES

- Bhargava, H.K., Bana, S.H., Zhang, Z., Loupos, P., Zantedeschi, D., Brandimarte, L., Choudhary, V., Li, J.F., 2026. Fighting fire with fire: infusing AI into peer review to sustain quality scholarship. Working paper, March 2, 2026.
- Chen, C., Kershaw, J.A. Jr., Weiskittel, A.R., McGarrigle, E., 2023. Can a multistage approach improve individual tree mortality predictions across the complex mixed-species and managed forests of eastern North America? *Forest Ecosystems* 10, 100086.
- Cieszewski, C.J., 2026. Artificial intelligence as a research instrument and an editorial burden: scientific judgment, forestry scientometrics, and academic incentives. *Math. Comput. For. Nat.-Res. Sci.* 18(1), 36–58.
- Cieszewski, C.J., Strub, M.R., 2009. The foundation of Mathematical and Computational Forestry & Natural-Resource Sciences (MCFNS). *Math. Comput. For. Nat.-Res. Sci.* 1, 41–46.
- Committee on Publication Ethics (COPE), 2019. Core practices and guidelines for peer review. <https://publicationethics.org/core-practices>.
- Cunningham, S., 2026. Claude Code 27: Research and publishing. *Causal Inference with Mixtape* [Substack blog], Feb. 2026. <https://causalinf.substack.com/p/claude-code-27-research-and-publishing>.
- Duijm, E., 2024. The coming apocalypse for scientific publishing. *The BS Detector* [Substack blog], Oct. 2024. <https://thebsdetecter.substack.com/p/the-coming-apocalypse-for-scientific>.
- Dutilleul, P., 1993. Modifying the t test for assessing the correlation between two spatial processes. *Biometrics* 49, 305–314.
- Frontiers Media, 2025. The State of Peer Review 2025: Global Survey of Researcher Attitudes and Practices. Frontiers Media S.A., Lausanne, Switzerland.
- García, J.A., Rodríguez-Sánchez, R., Fdez-Valdivia, J., 2015. The principal-agent problem in peer review. *J. Assoc. Inf. Sci. Technol.* 66, 297–308.
- Koppel, M., Schler, J., Argamon, S., 2009. Computational methods in authorship attribution. *J. Am. Soc. Inf. Sci. Technol.* 60, 9–26.
- Kustov, A., 2026. Academics need to wake up on AI: Ten theses for folks who haven't noticed the ground shifting under their feet. *Popular by Design* [Substack blog], Mar. 2, 2026. <https://www.popularbydesign.org/p/academics-need-to-wake-up-on-ai>.
- Li, W., Zhao, R., Sun, P.-T., Lee, J., Lu, O., Yuan, S., Sureshkumar, P., Chen, Y., Xiao, Z., Tang, Y.J., 2026. Large language models for peer review in biotechnology. *New Biotechnol.* 93, 218–225. <https://doi.org/10.1016/j.nbt.2026.03.007>.
- Osorio, F., Vallejos, R., 2024. SpatialPack: Computing the association between two spatial processes. R package version 0.4. <https://CRAN.R-project.org/package=SpatialPack>.
- Russell, M.B., Weiskittel, A.R., 2026. Improving peer review capacity and integrity in forestry journals. *Forest Ecosystems* 16, 100455. <https://doi.org/10.1016/j.fecs.2026.100455>.
- Tennant, J.P., Waldner, F., Jacques, D.C., Masuzzo, P., Collister, L.B., Hartgerink, C.H.J., 2016. The academic, economic and societal impacts of Open Access: an evidence-based review. *F1000Research* 5, 632.
- Walters, W.H., Wilder, E.I., 2023. Fabrication and errors in the bibliographic citations generated by ChatGPT. *Sci. Rep.* 13, 14045. <https://doi.org/10.1038/s41598-023-41032-5>.
- Weiskittel, A.R., 2015. Evaluating traditional peer-review processes and their alternatives: an opinionated discussion. *Math. Comput. For. Nat.-Res. Sci.* 7(2), 81–92.

APPENDIX A: SELF-ASSESSMENT OF THIS MANUSCRIPT USING THE MCFNS FRAMEWORK

The proposed guidelines are designed so that the B1–B12 rubric and six-module prompt library can be applied to any MCFNS submission. Applying them to the present paper is a consistency check and a transparency measure. The assessment was conducted by the author in his capacity as manuscript editor, which is an acknowledged conflict of interest; this appendix is the public record of that evaluation in lieu of an independent review.

The pre-revision manuscript scored 58.5/80 (73%), placing it in the Minor Revision band (65–84%); eight issues were identified across five dimensions, none of which triggered a critical override. Post-revision the score is 62/80 (77.5%), still in the Minor Revision band, and the editors accepted the paper once those revisions were complete. Table A1 gives the pre- and post-revision scoring matrix; Table A2 itemizes the issues and actions.

A.1 Scoring matrix: pre- and post-revision

Table A1: Self-assessment scoring matrix. Scale: 5 = Excellent, 4 = Good, 3 = Adequate, 2 = Below standard, 1 = Unacceptable. Wt. = dimension weight. Pre = score before revision; Post = score after revision; Δ = change. Both totals fall in the Minor Revision band (65–84%). No critical overrides triggered (no $\times 2$ dimension scored 1 at either stage). Weighted totals: Pre 58.5/80 (73%); Post 62/80 (77.5%).

Dim.	Evaluation dimension	Wt.	Pre	Post	Δ	Primary rationale
B1	Originality and novelty	$\times 1$	4	4	—	Four-premise retrospective against a named prior paper is novel for this literature; worked example is original demonstration content.
B2	Scientific significance	$\times 1$	4	4	—	Addresses an operational problem affecting all forestry journals; guidelines and worked example provide journal-level guidance.
B3	Methodological rigor	$\times 2$	3	4	+1	Three unsourced factual claims resolved (see Table A2, rows 1–2); platform closure dates retained as widely verifiable public record.
B4	Statistical appropriateness	$\times 2$	4	4	—	No original analyses; all attributed statistics correctly hedged.
B5	Mathematical/model rigor	$\times 2$	4	4	—	B1–B12 rubric internally consistent; no mathematical errors.
B6	Model validation*	$\times 2$	3	3	—	Rubric has no inter-rater reliability data; inherent first-edition limitation acknowledged; not a bar to publication.
B7	Data availability	$\times 1$	4	4	—	Worked example data in Table 1; prompt library in companion guidelines.
B8	Code / reproducibility	$\times 2$	4	4	—	Full prompt library (Modules 0–6b) published in companion document.
B9	Literature/citation quality	$\times 1$	3	4	+1	Reference list expanded; COPE, Koppel et al., Tennant et al., and Duijm among the additions (see Table A2, rows 3–4).
B10	Interpretation	$\times 1$	4	4	—	Four-premise retrospective consistent throughout; Section 2.4 hedging corrected (see Table A2, row 5).
B11	Writing clarity	$\times 0.5$	3	4	+1	Two mechanical errors corrected; five vocabulary signals removed (see Table A2, rows 6–8).
B12	Figures and tables	$\times 0.5$	4	4	—	Figures 1–2 and Table 1 appropriately labeled and captioned.
Weighted total			58.5	62	+3.5	
Percentage			73%	77.5%		
Rubric band			Minor Rev.	Minor Rev.		

* B6 (model validation, score 3 at both stages) reflects the absence of inter-rater reliability data for the rubric. The worked example in Section 6, scored through the same B1–B12 framework, produced a weighted total of 29/80 (36%) with three critical overrides and a Level 3 citation integrity finding, consistent with a Reject decision. These two cases span the rubric range from Minor Revision to Reject and are the first calibration evidence for the rubric; additional data points across future submission cycles will inform the v1.1 update. Calibration disagreements between the rubric and expert editorial judgment are the most informative category for that update.

Conflict of interest and integrity: The author of this paper is also the manuscript editor for MCFNS and lead developer of the guidelines being introduced. The scoring judgments reported here are not independent in the way external reviewer judgments would be; this appendix is offered as a demonstration of method and record of revision, not a substitute for independent review. Module 4 screening returned no Pattern A or B integrity anomalies. The accept decision was made by the editor, Chris J. Cieszewski, not by the author.

A.2 Issue disposition

Table A2: Issues identified during self-assessment and actions taken in revision. Dimension codes reference Table A1. All eight issues were resolved; none triggered a critical override.

Row	Dim.	Pre-revision finding	Post-revision action
1	B3, B9	45%/20% editorial board diversity statistic stated without citation.	Replaced with hedged geographic diversity claim citing Tennant et al. (2016); no specific figures asserted.
2	B9	“4% of papers usefully replicable” traced through Kustov (2026) to an unnamed source.	Duijm (2024) identified as primary source; sentence recast to flag the estimate as informal.
3	B9	No COPE citation despite recommending COPE escalation procedures.	COPE (2019) core practices added to reference list and cited in Section 2.4.
4	B9	No stylometrics primary literature supporting the double-blind authorship detection claim in Section 2.4.	Koppel et al. (2009) added; claim hedged as “likely deepened” and “raising the possibility.”
5	B10	Section 2.4 stated “This limitation has deepened” with more certainty than the available evidence supports.	Recast: “has likely deepened . . . raising the possibility that. . . The field has not yet formally assessed this risk.”
6	B11	Triple repetition of “one estimate puts” at line 291: a find-and-replace artifact that survived manual editing.	Corrected to a single clean clause with Duijm 2024 citation. Caught by Module 3 structural scan.
7	B11	Subsection header “premise (iii) and (iv)” grammatically incorrect; plural required.	Corrected to “premises (iii) and (iv).”
8	B11	Module 3 vocabulary signals: <i>well positioned</i> (abstract), <i>genuine scientific novelty</i> (Sec. 3.1), closing phrase “not yet organized” (vague), <i>substantially</i> repeated three times.	All four signals removed or varied; conclusion rewritten with a specific interoperability call to action.

A.3 Observations and limitations

Three of the eight findings survived repeated manual editing: a triple-phrase mechanical artifact (row 6), an unsourced statistic inserted as a draft placeholder (row 1), and a truncated citation trace (row 2). These correspond to three common error classes (mechanical artifact, incomplete draft task, truncated citation trace), each addressed by a different module in the prompt library. The self-assessment is consistent with the claim that the framework applies uniformly regardless of familiarity with the material, but a single case study does not establish that claim.

The only dimension revision cannot resolve is B6 (model validation, score 3 at both stages). The B1–B12 rubric has not been calibrated across multiple independent reviewers, and no single paper can supply that evidence; the v1.1 update will draw on cases accumulated across submission cycles. Editors and reviewers using these guidelines are specifically invited to record instances where the rubric produces counterintuitive results relative to their own expert judgment, since disagreement is more informative for calibration than agreement.